Serving the Needs of Model Intercomparison Projects and the IPCC

Karl Taylor

Program for Climate Model Diagnosis and Intercomparison
Lawrence Livermore National Laboratory

GO-ESSP Meeting

Didcot, England

6-8 June 2005

Characteristics of model intercomparison projects

- Original purposes:
  - → Identify common model errors
  - → Evaluate individual model performance relative to the group
  - → Identify robust responses
  - → Determine factors that are most important in explaining differences in model response
- Consequences for some projects: more open access to results
  - → Enables a more comprehensive analysis and scrutiny of model behavior by a wider community of experts
  - → Should speed advances in the understanding of model behavior

Ingredients important to successful model intercomparison projects include:

- Well defined experiment
- Sufficient participation
- Precisely defined set of model output
- Translation of all model output into a single format and structure, with sufficient metadata to perform analyses
  - → Reduces overall effort needed to analyze results
  - → Usually reduces the amount of data transferred to any individual analyst (because instead of files with many variables, but a single time slice, the file structure becomes a single variable, but many time-slices)

Model intercomparison projects have evolved

- Increasing amounts of data
- Stricter data requirements
- Task of translating model output to a common format shifting from coordinating groups to individual groups
- Wider community involvement
- Proliferating number of projects (AMIP, PMIP, CMIP, SMIP, C4MIP, CFMIP, IAEMIP, PILPS, NARCCAP, CCMVal, APE, …)

Growing data volume

- 1980's (e.g., FANGIO): a few time mean and global mean (or zonally averaged) quantities: ~0.0001 Gbyte
- Early 1990's (e.g., AMIP1, PMIP, CMIP1): modest collection of monthly mean 2-D fields: ~1 Gbyte
- Late 1990's (e.g., AMIP2): large collection of monthly mean and 6-hourly 2-D and 3-D fields: ~500 Gbytes
- Present (IPCC simulations): fairly comprehensive output from both ocean and atmospheric components; monthly, daily, & 3-hourly: ~25,000 Gbytes
- 2010: ~1000 Tbytes???

25 terabytes: How much data is that?

- A single latitude/longitude map at typical climate model resolution occupies about ~40 kilobytes.
- If you wanted to look at all 25 Tbytes in the form of these latitude/longitude plots, and if
  - → every 10 seconds you displayed another map
  - → you worked 50-hour weeks

  You could complete the task in about 600 years (or you could watch a "movie" of the output at 6 frames/sec for 10 years).
- If we divided up the task among the ~300 scientists registered to analyze the data, each scientist scientists (working 50-hour weeks), would have to look at a new plot every 10 seconds for 2 years.

Current IPCC model output archive statistics (5/27/05)

- Archive size: 22 Tbytes & 50,000 files
- Data distributed to users: 37 Tbytes & 190,000 files (mean file size ~ 200 Mbytes)
- Registered users: 344
- Mean download rate (over last 5 months) ~ 200 Gbytes/day, ~1000 files/day (peak ~500 Gbytes/day)

Challenges:

- How to distill from all this model output useful scientific information.
- We can already meet the challenge of collecting and distributing data on this scale, so no need to develop entirely new solutions, but:
  - → At each stage tools could be developed to improve efficiency
  - → Substantially more ambitious community modeling projects (>~200 Tbytes) may require a distributed database

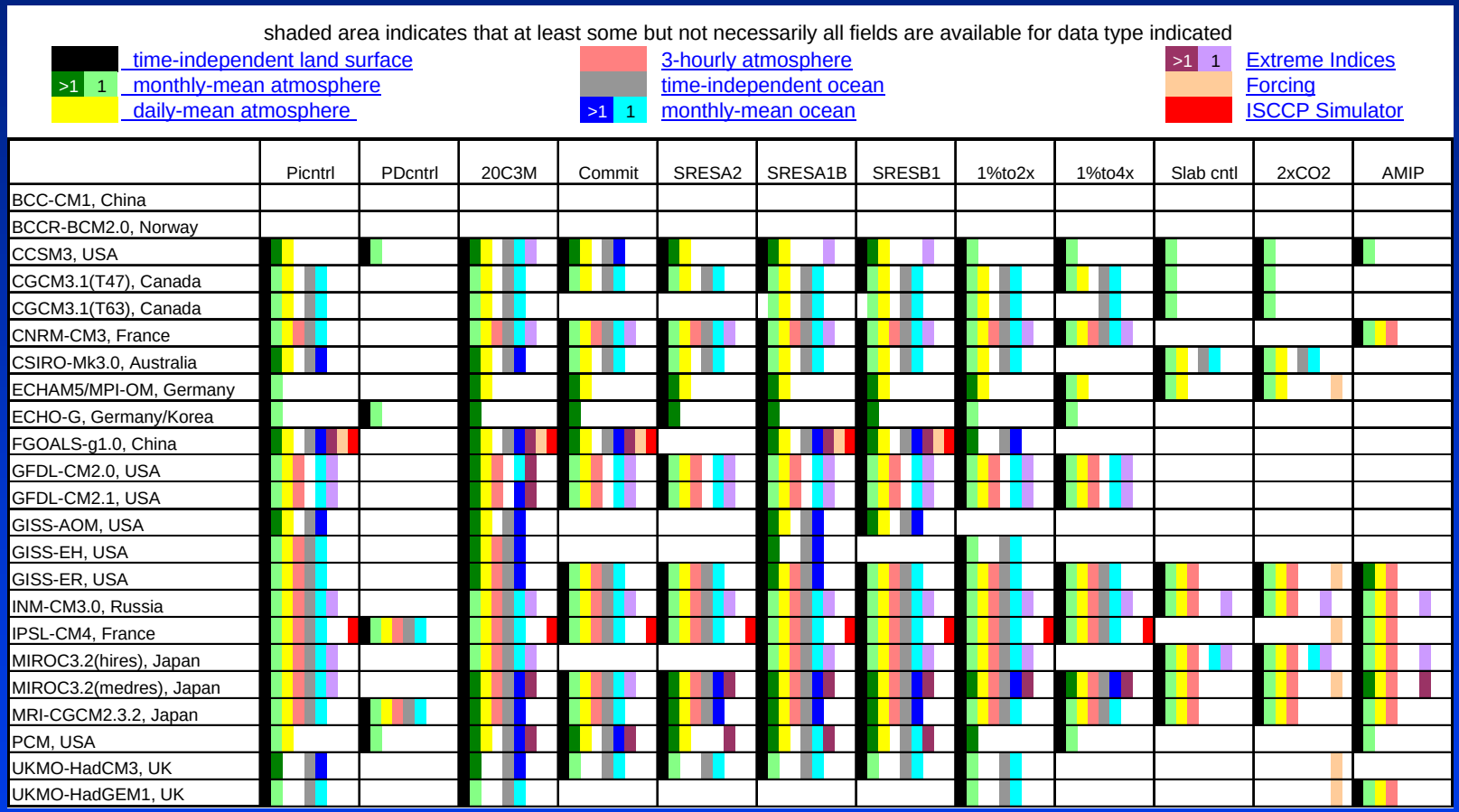
What are priorities for model intercomparison project?

- More reliable portable disks for transferring data from modeling centers to central distribution facility and automated backup.
- Alternatively, distributed database capability, but
  - → Some centers have severe security restrictions
  - → Requires more support at individual centers
- Extension/generalization of CMOR-like software for rewriting data.
- Further automation of data management procedures (all relatively small projects)

Useful incremental additions to current capabilities:

- Additional quality assurance software
- Tools to facilitate “publication” and cataloging of output on portals
- Automated updating of output availability/status pages
- Searchable errata pages
- Automated notification of users with updates tailored to their interests (new, withdrawn, replaced data)
- Creation of searchable model documentation
- Straight-forward enhancements of interactive searching tools (*not* sophisticated discovery tools)
- Scripts to perform common ftp download tasks

Data availability summary (as of 24 May 2005)



http://www-pcmdi.llnl.gov/ipcc/data_status_tables.htm

Sample errata entries (selected from 34)

Date	Model	Files	Description	Status
12/09/04	giss_model_e_r	All SRESA2 monthly files	The start time should be 2004-1-1, not 1901-1-1. The approximate time range is 2004 to 2100.	1/3/05: Updated files available.
1/3/05	miroc3_2_hires	/ipcc/1pctto2x/atm/yr/gsl/miroc3_2_hires/run1/gsl_A4.nc /ipcc/20c3m/atm/yr/gsl/miroc3_2_hires/run1/gsl_A4.nc /ipcc/picntrl/atm/yr/gsl/miroc3_2_hires/run1/gsl_A4.nc	Growing season length is wrong.	1/10/05: New files available.
1/3/05	mri_cgcm2_3_2a	/ipcc/*/atm/mo/<var>/mri_cgcm2_3_2a/*/<var>_A1*.nc where <var> = cl, clt, rldscs, rlutcs, rsdscs, rsuscs, rsutcs	Data withdrawn.	Files withdrawn. Awaiting replacement.
2/9/05	cnrm_cm3	All scenarios. Variables: tos, sic, sit, usi, vsi, wfo, stfbarot, zobt, so, thetao, rhopoto, uo, vo, wo, zmlo, sbl, hfsib, sltfsib	Land-sea mask is wrong (land area is too large).	Files withdrawn. Awaiting replacement
2/9/05	ukmo_hadcm3	Variable: ts	Data provided only on the ocean grid (N144).	2/9/05: Replacement data available on the atmosphere grid over land and ocean.
2/23/05	gfdl_cm2_0	Variable: sftlf (land area fraction)	Ocean cells had missing data flag (1.e20) instead of 0 values.	Replacement data available.
2/23/05	miroc3_2_medres	Variables: zosga, zostoga	Units should be '100m' ,not 'm'.	Files withdrawn. Awaiting replacement.

https://esg.llnl.gov:8443/about/errata.do

Bob Drach

What is ***not*** needed for model intercomparison projects:

- Server-side analysis or visualization tools (except possibly subsetting, concatinating, regridding capabilities)
- Sophisticated “discovery” capabilities
- Metadata describing each model’s experiment in exhaustive detail (but this may be useful at individual modeling centers)

Homogenizing requirements for data treatment in model intercomparison projects

In order to reduce effort required to participate in various model intercomparison projects:

- Archive CF-conforming netCDF files, structured similar to IPCC database.
- Start with IPCC output requirements and then add metadata as needed
- Start with IPCC standard output tables and then subset/add fields as needed
- Recommend use of CMOR and supply appropriate CMOR tables to facilitate conformance with requirements.
- When possible, use the IPCC variable names (for now)

Summary

- The standards set by the IPCC exercise are a solid base for building and embellishing
- Future progress may be more limited by antiquated analysis approaches and software tools, not on handling enormous data volumes
- A more uniform set of requirements and procedures in model intercomparison projects will stretch limited resources
- If a distributed database becomes necessary, rely on existing solutions (e.g., file sharing approaches like napster, bittorrent)


PCMDI